統計数理(2014)
第 62 巻 第 1 号 123–133

特集「疫学研究のデザインとデータ解析：最近の理論的展開と実践」
[総合報告]

分子疫学研究における統計的方法論の展開：がんサブタイプへの取り組み

口羽 文†

(受付 2013 年 9 月 3 日；改訂 12 月 22 日；採択 2014 年 1 月 6 日)

要　　旨

一般に，がんは臓器別に研究されることが多い．疫学研究においても，各部位のがんと曝露因子との研究から，これまでに数多くのリスク因子が同定されている．しかし，一方で，腫瘍の分子レベルでの特徴はさまざまであり，これらは固有の発症メカニズムを反映しているであろうと考えられる．よって，新しいリスク因子の同定や発症メカニズムに関するより深い洞察を得るために，分子データによって定義されるサブタイプ間の原因の違いを評価する研究は増えつつある．そこで，本稿では，腫瘍の分子レベルでの多様性を etiologic heterogeneity(原因の不均一性)の観点より評価するための近年の方法論の展開を紹介する．

キーワード：原因の不均一性，サブタイプ，競合リスク，クラスタリング，分子疫学，がんゲノム．

1. はじめに

がんの分子レベルでの特徴はさまざまである．このことは，がんの発生，増殖の生物学的メカニズムが多様であることを示しており，これらの分子的特徴の同定は，新しい治療薬ターゲットの発見や，個別化医療への手がかりを得ることに大きく貢献している(Harris and McCormick, 2010)．同じ部位のがんであっても，分子的特徴の違いによって予後や治療反応性が異なることを示す研究は，乳がん(Holmes et al., 2011; Perou et al., 2000; Sotiriou and Pusztai, 2009)や大腸がん(Karapetis et al., 2008; Morikawa et al., 2011)を含めここ 10 年でかなり発展している．一方，疫学研究においては，分子生物学的，あるいは臨床的に異なる性質を持つサブタイプは発症原因も異なるのか，つまり etiologic heterogeneity(原因の不均一性)に関する興味が高まっており，実際にサブタイプによってリスク因子が異なることを示唆する報告もなされている(Kuchiba et al., 2012; Ma et al., 2006; Schildkraut et al., 1997; Tamimi et al., 2012)．サブタイプ間による原因の違い，あるいは類似性を評価することによって，発症メカニズムのさらなる洞察を得ることに加え，新しいリスク因子の同定や将来の最適な研究デザインの提案につながることが期待されている．

分子マーカーの測定技術が進歩するにつれ，今後このような研究はますます増えていくことが予想されると同時に，この新しいデータに関連する方法論的研究も重要な分野になりつつある．本稿では，まず前向きコホート研究データを中心に，既知のサブタイプに対する最近の方法論の展開を紹介する．次に，原因の違いという観点よりサブタイプを同定するという新しい

†国立がん研究センター 生物統計部門：〒 104–0045 東京都中央区築地 5–1–1

概念的枠組みとその解析方法を紹介し，最後に今後の課題についてまとめる．

2. 既知のサブタイプと曝露因子との関連

コホート研究において，曝露因子とがん罹患との関連を Cox 比例ハザードモデルを用いて推定することを考える(Kalbfleisch and Prentice, 2002)．多くの疫学研究では，興味のある疾患は均一な一疾患として扱われるが，ここでは腫瘍マーカーデータによって，この疾患がさらに J 個のサブタイプに分類されるとする．興味のある課題は，「ある曝露因子の効果がサブタイプ間で異なるかどうか」である．たとえば，ある疾患が 2 つのサブタイプ A と B に分類できるとする．サブタイプ A に対する曝露効果とサブタイプ B に対する曝露効果が同じであれば，この曝露はこの疾患に共通する発症機序に寄与していると考えられる．一方で，サブタイプ間で曝露効果が異なる場合は，この曝露がある特定の発症機序により特異的に寄与している可能性を示唆していると考えられる．本稿では，曝露効果の違いを表すパラメータを原因の不均一性パラメータと呼ぶ．また，以降では，腫瘍のゲノム変化，病理学的所見，臨床的な指標などの腫瘍の特徴を示す指標はすべてマーカーと表現する．

2.1 競合リスク解析

J 個のサブタイプをそれぞれ異なるイベントと捉えれば，このデータは競合リスクデータと考えることができる．ここで，対象者は，J 個のサブタイプのうちのどれか 1 つを発症し得るとし，同時がんや繰り返しイベントは考えない．j 番目のサブタイプに対する cause-specific ハザード $\lambda_j(j=1,\ldots,J)$ は，比例ハザードモデルを用いて

$$\lambda_j(t|\boldsymbol{X}) = \lambda_{0j}(t)\exp(\boldsymbol{X}\boldsymbol{\beta}_j^T) \tag{2.1}$$

と表される(Kalbfleisch and Prentice, 2002; Prentice et al., 1978)．ここで，λ_{0j} は j 番目のサブタイプに対するベースラインハザード，$\boldsymbol{X}$ は共変量ベクトル，$\boldsymbol{\beta}_j$ は j 番目のサブタイプに対する $\boldsymbol{X}$ の対数ハザード比ベクトルである．$i(i=1,\ldots,n)$ 番目の対象者のイベントの種類を Y_i (0: 打ち切り，j: j 番目のサブタイプ)，追跡時間を t_i とすれば，部分尤度は以下のように書くことができる．

$$\begin{aligned} L(\boldsymbol{\beta}_1\ldots\boldsymbol{\beta}_J) &= \prod_{i=1}^{n}\prod_{j=1}^{J}\left\{\frac{\lambda_j(t_i|\boldsymbol{X}_i)}{\sum_{l=1}^{n} I(t_l\geq t_i)\lambda_j(t_i|\boldsymbol{X}_i)}\right\}^{I(Y_i=j)} \\ &= \prod_{j=1}^{J}\left[\prod_{i=1}^{n}\left\{\frac{\lambda_j(t_i|\boldsymbol{X}_i)}{\sum_{l=1}^{n} I(t_l\geq t_i)\lambda_j(t_i|\boldsymbol{X}_i)}\right\}^{I(Y_i=j)}\right] \\ &= \prod_{j=1}^{J}\left[\prod_{i=1}^{n}\left\{\frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta}_j^T)}{\sum_{l=1}^{n} I(t_l\geq t_i)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_j^T)}\right\}^{I(Y_i=j)}\right] = \prod_{j=1}^{J}\ell(\boldsymbol{\beta}_j) \end{aligned} \tag{2.2}$$

ここからわかるように，部分尤度 L は，各サブタイプに対する部分尤度 $\ell(\boldsymbol{\beta}_j),\ j=1,\ldots,J$ に分解することができる．ここで，$\ell(\boldsymbol{\beta}_j)$ は j 番目以外のサブタイプは発症時点で打ち切りとしたときに得られる部分尤度である．従って，すべての共変量の効果がサブタイプ間で異なることを前提とすれば，各サブタイプに対してそれぞれ Cox 回帰分析を行うことで $\boldsymbol{\beta}_j$ を推定することができる．

今，興味のあるパラメータは $\boldsymbol{\alpha}_{1j}=\boldsymbol{\beta}_j-\boldsymbol{\beta}_1(j=2,\ldots,J)$ である．ここで，$j=1$ はリファレンスサブタイプとする．ここでは，一つのモデルから $\boldsymbol{\alpha}_{1j}(j=2,\ldots,J)$ に対する推論を行う方法として Lunn and McNeil(1995)による data augmentation 法を紹介する．まず，augmented

表 1. Augmented データセットの例．id：対象者番号，time：追跡期間，cancer：サブタイプ番号(ただし打ち切りの場合は 0)，X：曝露変数(ここでは 0 か 1 をとる 2 値変数)，censor：打ち切り変数(0：打ち切り，1：発症)，$\mathrm{X}_{\mathrm{agm}j}$：$j$ 番目のサブタイプのための曝露変数，type：どのサブタイプに対するデータかを示す変数．

id	time	cancer	X	censor	$\mathrm{X}_{\mathrm{agm1}}$	$\mathrm{X}_{\mathrm{agm2}}$	...	$\mathrm{X}_{\mathrm{agm}J}$	type
1	20	1	1	1	1	0	...	0	1
1	20	1	1	0	0	1	...	0	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
1	20	1	1	0	0	0	...	1	J

データセットを作成する．ここで，各対象者のデータは J 回入力される．表 1 にその例を示す．この対象者($id = 1$)は $time = 20$ の時点で 1 番目のサブタイプに分類されるがんを発症している．他のサブタイプに対するデータ($type \neq 1$)に対しては，$time = 20$ で打ち切りとされる．新たな変数 $\boldsymbol{X}_{\mathrm{agm}j}(j = 1, \ldots, J)$ は，j 番目のサブタイプに対する曝露効果を評価するための変数であり，$type = j$ に対しては $\boldsymbol{X}_{\mathrm{agm}j} = \boldsymbol{X}$，それ以外では $\boldsymbol{X}_{\mathrm{agm}j} = 0$ をとる．

$\boldsymbol{X}_{\mathrm{agm}j}(j = 1, \ldots, J)$ を用いると，式(2.1)は以下のように書き換えることができる．

$$\begin{aligned}\lambda_j(t|\boldsymbol{X}) &= \lambda_{0j}(t)\exp(\boldsymbol{X}_{\mathrm{agm1}}\boldsymbol{\beta}_1^T + \boldsymbol{X}_{\mathrm{agm2}}\boldsymbol{\beta}_2^T + \cdots + \boldsymbol{X}_{\mathrm{agm}J}\boldsymbol{\beta}_J^T) \\ &= \lambda_{0j}(t)\exp(\boldsymbol{X}\boldsymbol{\beta}_1^T + \boldsymbol{X}_{\mathrm{agm2}}\boldsymbol{\alpha}_{12}^T + \cdots + \boldsymbol{X}_{\mathrm{agm}J}\boldsymbol{\alpha}_{1J}^T)\end{aligned} \tag{2.3}$$

また，もし一部の共変量 $\boldsymbol{W}$ に対してサブタイプにかかわらず共通効果 $\boldsymbol{\beta}^c$ を仮定したい場合，cause-specific ハザードは，

$$\begin{aligned}\lambda_j(t|\boldsymbol{X}) &= \lambda_{0j}(t)\exp(\boldsymbol{X}_{\mathrm{agm1}}\boldsymbol{\beta}_1^T + \boldsymbol{X}_{\mathrm{agm2}}\boldsymbol{\beta}_2^T + \cdots + \boldsymbol{X}_{\mathrm{agm}J}\boldsymbol{\beta}_J^T + \boldsymbol{W}\boldsymbol{\beta}^{cT}) \\ &= \lambda_{0j}(t)\exp(\boldsymbol{X}\boldsymbol{\beta}_1^T + \boldsymbol{X}_{\mathrm{agm2}}\boldsymbol{\alpha}_{12}^T + \cdots + \boldsymbol{X}_{\mathrm{agm}J}\boldsymbol{\alpha}_{1J}^T + \boldsymbol{W}\boldsymbol{\beta}^{cT})\end{aligned} \tag{2.4}$$

と書くことができる．Augmented データセットを用いて，サブタイプによる層別 Cox 回帰分析を行うことで，それぞれのサブタイプに対する曝露効果 $\boldsymbol{\beta}_j$，あるいは曝露効果(つまり，原因)の不均一性の指標である $\boldsymbol{\alpha}_{1j}$ を一つのモデルから同時に推定することができる．

次に，特に，J 個のサブタイプが複数のマーカーで定義されている場合を考える．例えば，大腸がんでは，マイクロサテライト不安定性(MSI: microsatellite instability)，CpG アイランドメチル化形質(CIMP: CpG island methylator phenotype)，*BRAF* 変異の有無は，腫瘍の特徴を表す代表的なマーカーとして知られている．それぞれのマーカーで定義されるサブタイプと喫煙との関連を検討した研究から，喫煙は MSI の高い大腸がんのリスクは増加させるが，MSI の低い大腸がんとは関連がみられないことが示唆されている(Chia et al., 2006; Limsui et al., 2010; Poynter et al., 2009; Samowitz et al., 2006)．同様に，喫煙は CIMP の高い大腸がんのリスクは増加させるが，CIMP の低いがんとは関連がないこと(Limsui et al., 2010; Samowitz et al., 2006)，また，*BRAF* が変異している大腸がんのリスクは増加させるが，*BRAF* が変異していない大腸がんとの関連はみられないこと(Limsui et al., 2010; Rozek et al., 2010; Samowitz et al., 2006)がそれぞれ示唆されている．一方で，腫瘍マーカー間の関係を検討した研究より，CIMP の高い大腸がんは MSI も高い傾向があり，また，さらに *BRAF* 変異の頻度も高いことが示されている(Hughes et al., 2012; Tanaka et al., 2010)．今のところ，喫煙がこれらのどの分子変化とも直接関連しているのか，あるいは，どれか一つの分子変化のみと関連していて，他の分子サブタイプとの関連はマーカー間の相関により間接的に見られているだけなのかは明らかではない．この例のように，多くの腫瘍マーカーは，相互に複雑に関連している．よって，他のマー

カーの影響を考慮した下で，曝露因子の効果の違いを反映している重要な腫瘍マーカー，あるいは腫瘍マーカーの組み合わせを探索することが課題となる．

今，K 個のマーカーが測定されているものとする．k 番目のマーカーの値を m_k とし，このマーカーによって疾患は M_k 個のカテゴリに分類できるとすると，潜在的には $M_1 \times M_2 \times \cdots \times M_K$ 個のサブタイプが存在することになる．ここで，解析上の問題が 2 点あげられる．まず，組み合わせにより定義されるサブタイプの数は潜在的に大きくなり，それに伴い曝露効果パラメータの数も多くなること，次に，マーカーは変異の有無のように離散変数のものもあれば，mRNA 発現量のように連続量で得られるものもあり，型の異なるアウトカム変数を同時にモデル化しなければならないことである．これらの問題を解決するために，Chatterjee et al.(2010) は β_j に対する 2 段階モデルを提案した．ここでは，簡単のため，β_j はある一つの曝露変数に対する回帰係数とし，スカラーとする．興味のある疾患が，$Y = j = (m_1, m_2, \ldots, m_K)$ と K の特徴で表現できることを踏まえると，式(2.1)の β_j に以下のようなモデルを考えることができる．

$$\beta_{(m_1,m_2,\ldots,m_K)} = \theta^{(0)} + \sum_{k=1}^{K} \theta^{(1)}_{k(m_k)} + \sum_{k=1}^{K} \sum_{k' \geq k}^{K} \theta^{(2)}_{kk'(m_k,m_{k'})} + \cdots + \theta^{(K)}_{12\cdots K(m_1,\ldots,m_K)} \tag{2.5}$$

ここで，$\exp(\theta^{(0)})$ はリファレンスサブタイプ($Y = (0, 0, \ldots, 0)$)に対するハザード比，$\theta^{(1)}_{k(m_k)}$ は k 番目のマーカーデータによるサブタイプ間の曝露効果の違いを示す指標であり，サブタイプ $Y = (0, 0, \ldots, 0)$ と $Y = (0, \ldots, m_k, \ldots, 0)$ 間の(1 次の)不均一性パラメータである．$\theta^{(2)}_{kk'(m_k,m_{k'})}$ は $Y = (0, 0, \ldots, 0)$ と任意の二つのマーカーで定義されるサブタイプ $Y = (0, \ldots, m_k, m_{k'}, \ldots, 0)$ 間の不均一性パラメータ(2 次の不均一性パラメータ)であり，同様に K 次の不均一性パラメータまで考えることができる．式(2.5)は，β_j を各マーカー成分への曝露効果に分解したものと考えられる．2 次以降の不均一性パラメータを 0 と仮定すれば，モデル(2.5)は，

$$\beta_{(m_1,m_2,\ldots,m_K)} = \theta^{(0)} + \sum_{k=1}^{K} \theta^{(1)}_{k(m_k)} \tag{2.6}$$

と縮小でき，$\beta_{(m_1,\ldots,m_k,m_{k+1},\ldots,m_K)} - \beta_{(m_1,\ldots,m^*_k,m_{k+1},\ldots,m_K)} = \theta^{(1)}_{k(m_k)} - \theta^{(1)}_{k(m^*_k)}$ は，k 番目以外のマーカーによる腫瘍の形質は同じであるという条件の下で，k 番目のマーカーで定義されるサブタイプ間での曝露効果の違い(つまり，不均一性)の程度を表す．また，$K = 1$ の場合，式(2.6)によるモデルは式(2.3)と同じになる．さらに，マーカーデータが順序変数，あるいは連続量として得られる場合，適切なスコア $s^{(k)}_{m_k}$ を与えることで

$$\theta^{(1)}_{k(m_k)} = \theta^{(1)}_k s^{(k)}_{m_k}, \quad k = 1, \ldots, M_k \tag{2.7}$$

と表現できる．ここで，$s^{(k)}_1 = 0$, $s^{(k)}_1 \leq s^{(k)}_2 \leq s^{(k)}_3 \leq \cdots \leq s^{(k)}_{M_k}$ とする．$\theta^{(1)}_k = 0$ のとき，k 番目のマーカーによるサブタイプ間では曝露効果に違いがないことを意味する．モデル(2.6)に，マーカー間の交互作用項を含めれば，あるマーカーで定義されるサブタイプ間の不均一性の程度が他のマーカーによって変化するかどうかを評価することができる．Rosner et al.(2013)も同様のモデルを提案している．さらに，彼らはモデル(2.6)を用いて，マーカー調整ハザード比，

$$HR_{adj}(X|m_k) = \exp\left(\theta^{(0)} + \theta^{(1)}_{k(m_k)} + \sum_{k'=1(k' \neq k)}^{K} \sum_{q \in M_{k'}} \theta^{(1)}_{k'(m_{k'})} \Pr(m_{k'} = q)\right) \tag{2.8}$$

を提案した．これは，他のマーカーを調整した下での，k 番目のマーカーが m_k であるサブタイプに対するハザード比と解釈できる．

もう一つのアプローチとして，メタ回帰分析による方法を紹介する(Kuchiba et al., 2014)．こ

こでは，まずモデル(2.1)を用いて $\beta_j (j = 1, \ldots, J; J = M_1 \times M_2 \times \cdots \times M_K)$ を推定する．m_{kj} を，j 番目のサブタイプの k 番目のマーカーの水準とすると，$\hat{\beta}_j$ に対して，メタ回帰モデル

$$\hat{\beta}_j = \gamma_0 + \sum_{k=1}^{K} \gamma_k m_{kj} + e_j, \mathrm{var}(e_j) = \widehat{\mathrm{var}}(\hat{\beta}_j) \tag{2.9}$$

を考えることができる．ここで，γ_k は k 番目のマーカーで定義されるサブタイプ間で曝露効果が異なるかどうかの指標である．当然，マーカー同士の交互作用項も含めることができる．本質的には，モデル(2.9)より推定される γ_k は，Chatterjee et al.(2010)や Rosner et al. (2013)より提案された方法による不均一性パラメータ(ここではモデル(2.6)の $\theta^{(1)}$)と同等である．彼らの方法では，パラメータ β_j に階層的なモデルを構築し，不均一性パラメータを一つのモデルから推定するのに対し，メタ回帰分析は，先に β_j を推定し，その後で不均一性パラメータを推定する 2 ステップアプローチと解釈することができる．

2.2 発症ケースのみを用いた解析

ここで，少しだけ発症ケースのみを用いた解析方法について触れる．各サブタイプに対する曝露効果の違い，つまり $\boldsymbol{\alpha}_{1j} = \boldsymbol{\beta}_j - \boldsymbol{\beta}_1 (j = 2, \ldots, J)$ は，非罹患者のデータがなくても推定することができる．疾患ケースのみを用いるケース・ケースデザインは遺伝子-環境交互作用を評価できる研究デザインとして方法論的研究も進んでいるが，不均一性パラメータの推定にも有用である(Begg and Zhang, 1994)．式(2.1)より，リファレンスサブタイプに対するハザードとサブタイプ j に対するハザードの比は，

$$\begin{aligned} \frac{\lambda_j(t|\boldsymbol{X})}{\lambda_1(t|\boldsymbol{X})} &= \lim_{\Delta t \to 0} \frac{p(t \le T < t + \Delta t, j^{th} subtype | T \ge t, \boldsymbol{X} = \boldsymbol{0})}{p(t \le T < t + \Delta t, 1^{st} subtype | T \ge t, \boldsymbol{X} = \boldsymbol{0})} \exp(\boldsymbol{X}(\boldsymbol{\beta}_j^T - \boldsymbol{\beta}_1^T)) \\ &= \frac{p(j^{th} subtype | T = t, \boldsymbol{X} = \boldsymbol{0})}{p(1^{st} subtype | T = t, \boldsymbol{X} = \boldsymbol{0})} \exp(\boldsymbol{X}(\boldsymbol{\beta}_j^T - \boldsymbol{\beta}_1^T)) \\ &= \frac{p(j^{th} subtype | T = t, \boldsymbol{X} = \boldsymbol{0})}{p(1^{st} subtype | T = t, \boldsymbol{X} = \boldsymbol{0})} \exp(\boldsymbol{X} \alpha_{1j}^T) \end{aligned} \tag{2.10}$$

となり，これは時点 T での多項ロジスティックモデルである．つまり，リファレンスサブタイプ群と他のサブタイプ群とを比較することで直接不均一性パラメータ α を推定できる．ただし，サブタイプ特有の曝露効果 β_j は推定できない．また，モデルパラメータが不均一性パラメータであることを考えれば，ある交絡因子に対して，サブタイプ間で効果が変わらないことが仮定できる場合，そのような因子はモデルに含める必要がない．ケース・ケースデザインを用いた場合のパラメータ α に対する推定効率に関しては，今後の検討課題である．

3. サブタイプの探索

次世代 DNA シークエンス技術に代表されるように，近年の分子データ測定技術の進歩は目覚ましく，より複雑な多次元オミクスデータによって腫瘍の特徴，多様性を把握することができる．一方で，曝露因子に関しても，ゲノムワイド関連研究のように網羅的に遺伝的なリスク因子が探索されている．よって，今後は曝露，アウトカムともにゲノムワイドな多変量データが得られ，このようなデータを疫学研究の中でどう解析，解釈していくかは重要なテーマになると考えられる．前章では，「既知のサブタイプ間での曝露因子の効果の違い」をモデル化する方法について述べたが，ここでは，Begg et al.(2013)により提案された「原因が異なるようにがんの分子サブタイプを定義する」という新しい概念とその枠組みを紹介する．

腫瘍マーカーデータが多次元になるほど，それらの類似性から定義できるサブタイプ分類は

一通りではない．ここでは，複数考えられるサブタイプ分類のうち，発症原因が似ているか異なるか(不均一性)という点より最適である分類を同定することが目的である．また，特定の曝露に注目するのではなく，リスク因子の集合で定義される全体的なリスク(リスクプロファイル)で「原因」を定義する．たとえば，各対象者のあるサブタイプに対する発症確率は，そのサブタイプに対するリスクプロファイルと考えることができる．このリスクプロファイルと腫瘍マーカーデータを双方向から用いながらサブタイプを同定していく．

3.1 集団におけるリスクの不均一性

i 番目の対象者の疾患リスクプロファイルを $r_i (i = 1, \ldots, n)$ とすると，対象集団でのリスクプロファイルの平均，分散はそれぞれ $\mu = n^{-1} \sum_{i=1}^{n} r_i, \nu = n^{-1} \sum_{i=1}^{n} r_i^2 - \mu^2$ となる．ここでは，リスクプロファイルのばらつき(つまり，不均一性)の指標として，スケール不変である変動係数 $\kappa = \sqrt{\nu}/\mu$ を用いる．

3.2 原因の不均一性

この疾患が 2 つのサブタイプ A, B に分類できるとする．サブタイプ A に対するリスクが高い(低い)対象者は，サブタイプ B に対しても高い(低い)リスクを持つ傾向にあるとき，この 2 つのサブタイプの原因は似ているということができる．つまり，各サブタイプに対するリスクプロファイル間の相関は，原因の不均一性を反映しているといえる．r_{Ai}, r_{Bi} はそれぞれ i 番目の対象者のサブタイプ A に対するリスクプロファイルとサブタイプ B に対するリスクプロファイルとする．これらは互いに排反とし，$r_i = r_{Ai} + r_{Bi}$ とする．つまり，対象者 i が興味のある疾患を発症する確率は，各サブタイプを発症する確率に分解できる．

集団において，サブタイプ A, B に対するリスクプロファイルの平均，分散は，それぞれ $\mu_A = n^{-1} \sum_{i=1}^{n} r_{Ai}, \mu_B = n^{-1} \sum_{i=1}^{n} r_{Bi}, \nu_A = n^{-1} \sum_{i=1}^{n} r_{Ai}^2 - \mu_A^2, \nu_B = n^{-1} \sum_{i=1}^{n} r_{Bi}^2 - \mu_B^2$ である．また，各サブタイプに対するリスクプロファイルの変動係数は，$\kappa_A = \sqrt{\nu_A}/\mu_A, \kappa_B = \sqrt{\nu_B}/\mu_B$ となる．さらに，リスクプロファイル間の共分散に対する標準化指標として，$\kappa_{AB} = \frac{n^{-1} \sum r_{Ai} r_{Bi} - \mu_A \mu_B}{\mu_A \mu_B}$ を考える．サブタイプ A とサブタイプ B 間の発症原因の違いが大きいほど，κ_{AB} は小さくなる．すなわち，κ_{AB} の大きさは不均一性の程度と逆の方向に変化する．また，リスクプロファイル間の相関は $\rho_{AB} = \kappa_{AB}/\kappa_A \kappa_B$ と表すことができる．

次に，サブタイプ数 $J > 2$ とする．ここでは，簡単のため 3 つのサブタイプ A, B, C で説明を行うが，サブタイプ数がさらに多くなった場合への一般化は容易である．興味のある疾患全体に対するリスクプロファイルの変動係数の二乗は以下のように分解できる．

$$\kappa^2 = \pi_A^2 \kappa_A^2 + \pi_B^2 \kappa_B^2 + \pi_C^2 \kappa_C^2 + 2\pi_A \pi_B \kappa_{AB} + 2\pi_A \pi_C \kappa_{AC} + 2\pi_B \pi_C \kappa_{BC} \tag{3.1}$$

ここで，$\pi_j = \mu_j/(\mu_A + \mu_B + \mu_C), j = A, B, C$ はこの集団におけるサブタイプ j の割合である．式(3.1)より，この疾患をどのようにサブタイプに分類したとしても κ^2 は一定であることに注目する．κ^2 はこの疾患全体に対するリスクプロファイルのばらつきであり，各サブタイプに対するリスクプロファイルのばらつきの集合体である．したがって，より原因の違いが顕著になるように(つまり，原因の不均一性が高くなるように)サブタイプを分類するということは，すなわち，3 つの分散 ν_A, ν_B, ν_C の合計が大きくなるようにサブタイプに分類することと同義である．ここで，

$$D = (\pi_A \kappa_A^2 + \pi_B \kappa_B^2 + \pi_C \kappa_C^2) - \kappa^2 \tag{3.2}$$

を定義する．これは，各サブタイプで説明できるリスクプロファイルのばらつきの平均と全体でのリスクプロファイルのばらつきの差である．通常，$D \geq 0$ である．すべてのサブタイプの

リスクプロファイルが集団の中で完全に相関している，つまり，どのサブタイプに対するリスクも同じ場合，$D = 0$(no heterogeneity)となる．式(3.2)はさらに

$$D = \pi_A\pi_B(\kappa_A^2 + \kappa_B^2 - 2\kappa_{AB}) + \pi_A\pi_C(\kappa_A^2 + \kappa_C^2 - 2\kappa_{AC}) + \pi_B\pi_C(\kappa_B^2 + \kappa_C^2 - 2\kappa_{BC}) \tag{3.3}$$

と書き換えることができる．式(3.3)をみると，リスクプロファイル間の共分散が小さいほど，D が大きくなることがよりわかりやすい．

3.3 解析方法

ここでの目標は，前章のようになんらかの事前情報を基にサブタイプを特定し，それらの間で原因因子が異なるかどうかを評価することではなく，発症原因が互いに異なるようにサブタイプに分類することである．つまり，D を最大にするようなサブタイプ分類方法を構築することが目的である．以下では，ケース・コントロール研究データを例にリスクプロファイルの推定とサブタイプのクラスタリングをどう組み合わせるかについて説明する．

3.3.1 リスクプロファイルの推定

ケース・コントロール研究データを用いて，ロジスティック回帰モデル $\log(p_i/(1-p_i)) = \boldsymbol{X}_i\boldsymbol{\eta}^T$ から，リスク因子 $\boldsymbol{X}$ の回帰係数ベクトル $\boldsymbol{\eta}$ を推定し，それらを用いて対象者 i の疾患リスクプロファイルを $\hat{p}_i = \exp(\boldsymbol{X}_i\hat{\boldsymbol{\eta}}^T)/(1+\exp(\boldsymbol{X}_i\hat{\boldsymbol{\eta}}^T))$ と推定することができる．ただし，ケース・コントロール研究から絶対リスクを推定することは困難なため，ここではリスクプロファイルの表記法に前節で用いた r ではなく p を用いる．今推定したい κ は，対象集団でのリスクプロファイルのばらつきであるため，$\boldsymbol{\eta}$ はケース，コントロールの両集団を用いて推定されるが，κ はコントロール集団のみの $\hat{p}_i$ を用いて推定することに注意する．コントロールの対象者を $i(i = 1, \ldots, n_0)$ とすれば，$\hat{\kappa}^2 = n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_i^2/(n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_i)^2 - 1$ と求めることができる．ここで，何らかの方法によりこの疾患が J 個のサブタイプに分類されているとする．対象者 i のリスクプロファイルは，各サブタイプに対するリスクプロファイルの和，$\hat{p}_i = \sum_{j=1}^{J}\hat{p}_{ji}$，であるので，各サブタイプに対するリスクプロファイルは多項ロジスティック回帰モデルを用いて推定することができ，それぞれの変動係数も同様に計算できる．つまり，各サブタイプに対するリスクプロファイルは，$\hat{p}_{ji} = \exp(\boldsymbol{X}_i\hat{\delta}_j^T)/(1 + \sum_{j=1}^{J}\exp(\boldsymbol{X}_i\hat{\delta}_j^T))$（ただし，$\delta_j$ はサブタイプ j に対するリスク因子の回帰係数)と推定することができ，それらを用いて，$\hat{\kappa}_j^2 = n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_{ji}^2/(n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_{ji})^2 - 1$，$\hat{\kappa}_{jj'}^2 = n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_{ji}\hat{p}_{j'i}/((n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_{ji})(n_0^{-1}\sum_{i=1}^{n_0}\hat{p}_{j'i})) - 1$ と計算される．これらの推定値を用いて腫瘍マーカーで定義された候補サブタイプ分類に対して，D を求めることができる．

3.3.2 最適サブタイプの同定

原因の違いが互いに最も顕著になるようにサブタイプ分類を行うためには，腫瘍間の特徴の違い，あるいは類似性とリスクプロファイルの不均一性あるいは類似性を結びつける必要がある．まず，適当な方法を用いて腫瘍マーカーに基づくサブタイプ分類を行う．このときにできたサブタイプの集合を U とする．サブタイプ分類の候補を $U_1, U_2, \ldots$ と作成し，この中で D が最大になるサブタイプ分類を最適な分類法とする．ここでは，K-means クラスタリング(Hartigan and Wong, 1979)を用いてサブタイプ分類を行う．この方法では，クラスタ間の腫瘍マーカーデータのばらつきに対して，クラスタ内の腫瘍マーカーデータのばらつきが最小となるようにクラスタが作成される．このクラスタは腫瘍マーカーデータのみで作られる，ということを強調しておく．対象者 i の腫瘍マーカーデータを Y_i とする．今，J 個のクラスタがあるとし，それぞれのクラスタ平均を $\varphi_j(j = 1, \ldots, J)$，全体での平均を φ で表す．また各対象者がどのクラスタに属するかを示す指示変数を d_{ij} とし，対象者 i がクラスタ j に属する場合は $d_{ij} = 1$，その他のクラスタに属する場合は $d_{ij} = 0$ をとる．K-means 法ではクラスタ間の相違は，$G = (S_T - S_I)/S_T$

で定義される．ただし，$S_I = \sum_{i,d_{ij}=1}\{(Y_i - \hat{\varphi}_j)(Y_i - \hat{\varphi}_j)'\}$，$S_T = \sum_i \{(Y_i - \hat{\varphi})(Y_i - \hat{\varphi})'\}$ である．K-means 法では，局所的最適解を得ることはできるが，初期値に大きく依存することが知られている．よって，クラスタ数を固定した下で初期値を変えることで，局所的に最適な候補クラスタ分類を $U_1, U_2, \ldots$ と何通りか作成し，それぞれに対して 3.1.1 より D を求め，D が最大のクラスタ分類を同定する．

この方法は，腫瘍マーカーデータのばらつきを示す G がリスクプロファイルのばらつき κ を反映することを前提としている．この点において，最適な G，κ の指標はまだ改善の余地があるかもしれない．また，最適なクラスタ数の決定に関しても今後の課題であるといえる．

4. まとめと今後の課題

がんにはそれぞれ個性があり，その多様性には発症機序の違いが反映されているだろうという考えは，それほど新しいものではないかもしれないが，膨大な分子データが入手しやすくなり，腫瘍の多様性をより詳細に把握できるようになるにつれ，疾患の多様性を「原因」から説明しようという試みはますます広がっていくものと考えられる．疫学研究において，「原因の不均一性」に対する考え方は整理され始めたばかりであるが，本稿では，大きく二つのアプローチを紹介した．まず第 2 章で，サブタイプ特異的なリスク因子を探索する方法を取り上げた．これは，いままでの疫学研究的方法の素直な拡張とみることができる．このアプローチは，比較的少数の既知のサブタイプに興味があるときにより有用である．一方，第 3 章では，腫瘍のプロファイル，リスクプロファイルともに多変量で得られる場合の考え方の一つを紹介した．オミクスプロファイルデータの到来を考えれば，このような包括的な方法の必要性は一層高まっていくことが予想される．

以下では分子データに関連するいくつかの課題についてまとめたい．既存コホート研究データの利用は，コスト効率の点で有用であるが，コホート内のすべての疾患発症者から腫瘍組織を収集することはほぼ不可能である．全米規模の大規模コホート研究である Nurses' Health Study や Health Professionals Follow-Up Study では，多くのがん種で腫瘍組織の収集に力を入れているが，たとえば大腸がんでの組織回収率は 60–70%程度である．マーカーデータが得られるケースのみを用いた解析では，検出力の低下を招くだけでなく，バイアスがかかる可能性がある．より多くの腫瘍組織を回収するための体制を整えることはもちろんだが，このような欠測データに対する解析的な取り組みも必要であると考えられる．

マーカーデータの質の管理も大きな課題である．測定誤差の原因となる要因の特定や，これらの不均一性パラメータへの影響，あるいはクラスタ分類への影響はこれから検討される必要がある．さらに，今後は測定方法の標準化や，研究結果の再現性，妥当性を評価する指針や枠組みを構築することが必要であると考えられる．

最後に，オミクスデータのような膨大なデータの入手が現実的になるにつれて，より効率的な研究デザインや，さまざまなタイプのデータを同時に組み入れることができる統計学的方法論の開発は重要な分野になると考えられる．また，どの視点からデータを捉えるか，つまり，研究の課題設定が非常に重要になり，多分野にまたがる研究者の協力体制は必須である．

謝　辞

貴重なコメントを下さった査読者に深く感謝申し上げます．

参 考 文 献

Begg, C. B. and Zhang, Z. F. (1994). Statistical analysis of molecular epidemiology studies employing case-series, *Cancer Epidemiology Biomarkers & Prevention*, **3**, 173–175.

Begg, C. B., Zabor, E. C., Bernstein, J. L., Bernstein, L., Press, M. F. and Seshan, V. E. (2013). A conceptual and methodological framework for investigating etiologic heterogeneity, *Statistics in Medicine*(to appear).

Chatterjee, N., Sinha, S., Diver, W. R. and Feigelson, H. S. (2010). Analysis of cohort studies with multivariate and partially observed disease classification data, *Biometrika*, **97**, 683–698.

Chia, V. M., Newcomb, P. A., Bigler, J., Morimoto, L. M., Thibodeau, S. N. and Potter, J. D. (2006). Risk of microsatellite-unstable colorectal cancer is associated jointly with smoking and non-steroidal anti-inflammatory drug use, *Cancer Research*, **66**, 6877–6883.

Harris, T. J. and McCormick, F. (2010). The molecular pathology of cancer, *Nature Reviews*, *Clinical Oncology*, **7**, 251–265.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm, *Journal of the Royal Statistical Society*, *Series C*(*Applied Statistics*), **28**, 100–108.

Holmes, M. D., Chen, W. Y., Schnitt, S. J., Collins, L., Colditz, G. A., Hankinson, S. E. and Tamimi, R. M. (2011). COX-2 expression predicts worse breast cancer prognosis and does not modify the association with aspirin, *Breast Cancer Research and Treatment*, **130**, 657–662.

Hughes, L. A., Khalid-De Bakker, C. A., Smits, K. M., Van Den Brandt, P. A., Jonkers, D., Ahuja, N., Herman, J. G., Weijenberg, M. P. and Van Engeland, M. (2012). The CpG island methylator phenotype in colorectal cancer: Progress and problems, *Biochimica et Biophysica Acta*, **1825**, 77–85.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New Jersey.

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., O'callaghan, C. J., Tu, D., Tebbutt, N. C., Simes, R. J., Chalchal, H., Shapiro, J. D., Robitaille, S., Price, T. J., Shepherd, L., Au, H. J., Langer, C., Moore, M. J. and Zalcberg, J. R. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer, *The New England Journal of Medicine*, **359**, 1757–1765.

Kuchiba, A., Morikawa, T., Yamauchi, M., Imamura, Y., Liao, X., Chan, A. T., Meyerhardt, J. A., Giovannucci, E., Fuchs, C. S. and Ogino, S. (2012). Body mass index and risk of colorectal cancer according to fatty acid synthase expression in the nurses' health study, *Journal of the National Cancer Institute*, **104**, 415–420.

Kuchiba, A., Wang, M., Spiegelman, D. (2014). Two-stage approach for identifying tumor subtypes associated with an exposure, Abstract # 312835, 2014 Joint Statistical Meetings, August 2–7 at the Boston Convention and Exhibition Center(accepted).

Limsui, D., Vierkant, R. A., Tillmans, L. S., Wang, A. H., Weisenberger, D. J., Laird, P. W., Lynch, C. F., Anderson, K. E., French, A. J., Haile, R. W., Harnack, L. J., Potter, J. D., Slager, S. L., Smyrk, T. C., Thibodeau, S. N., Cerhan, J. R. and Limburg, P. J. (2010). Cigarette smoking and colorectal cancer risk by molecularly defined subtypes, *Journal of the National Cancer Institute*, **102**, 1012–1022.

Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks, *Biometrics*, **51**, 524–532.

Ma, H., Bernstein, L., Pike, M. C. and Ursin, G. (2006). Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: A meta-analysis of epidemiological studies, *Breast Cancer Research*, **8**, R43.

Morikawa, T., Kuchiba, A., Yamauchi, M., Meyerhardt, J. A., Shima, K., Nosho, K., Chan, A. T., Giovannucci, E., Fuchs, C. S. and Ogino, S. (2011). Association of CTNNB1(beta-catenin) alterations, body mass index, and physical activity with survival in patients with colorectal

cancer, *JAMA*(*The Journal of American Medical Association*), **305**, 1685–1694.

Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O. and Botstein, D. (2000). Molecular portraits of human breast tumours, *Nature*, **406**, 747–752.

Poynter, J. N., Haile, R. W., Siegmund, K. D., Campbell, P. T., Figueiredo, J. C., Limburg, P., Young, J., Le Marchand, L., Potter, J. D., Cotterchio, M., Casey, G., Hopper, J. L., Jenkins, M. A., Thibodeau, S. N., Newcomb, P. A. and Baron, J. A. (2009). Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status, *Cancer Epidemiology, Biomarkers & Prevention*, **18**, 2745–2750.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Jr., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics*, **34**, 541–554.

Rosner, B., Glynn, R. J., Tamimi, R. M., Chen, W. Y., Colditz, G. A., Willett, W. C. and Hankinson, S. E. (2013). Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers, *American Journal of Epidemiology*, **178**, 296–308.

Rozek, L. S., Herron, C. M., Greenson, J. K., Moreno, V., Capella, G., Rennert, G. and Gruber, S. B. (2010). Smoking, gender, and ethnicity predict somatic BRAF mutations in colorectal cancer, *Cancer Epidemiology, Biomarkers & Prevention*, **19**, 838–843.

Samowitz, W. S., Albertsen, H., Sweeney, C., Herrick, J., Caan, B. J., Anderson, K. E., Wolff, R. K. and Slattery, M. L. (2006). Association of smoking, CpG island methylator phenotype, and V600E BRAF mutations in colon cancer, *Journal of the National Cancer Institute*, **98**, 1731–1738.

Schildkraut, J. M., Bastos, E. and Berchuck, A. (1997). Relationship between lifetime ovulatory cycles and overexpression of mutant p53 in epithelial ovarian cancer, *Journal of the National Cancer Institute*, **89**, 932–938.

Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer, *The New England Journal of Medicine*, **360**, 790–800.

Tamimi, R. M., Colditz, G. A., Hazra, A., Baer, H. J., Hankinson, S. E., Rosner, B., Marotti, J., Connolly, J. L., Schnitt, S. J. and Collins, L. C. (2012). Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer, *Breast Cancer Research and Treatment*, **131**, 159–167.

Tanaka, N., Huttenhower, C., Nosho, K., Baba, Y., Shima, K., Quackenbush, J., Haigis, K. M., Giovannucci, E., Fuchs, C. S. and Ogino, S. (2010). Novel application of structural equation modeling to correlation structure analysis of CpG island methylation in colorectal cancer, *The American Journal of Pathology*, **177**, 2731–2740.

Proceedings of the Institute of Statistical Mathematics Vol. 62, No. 1, 123–133 (2014) 133

Integration of Tumor Molecular Features into Epidemiologic Studies for Assessing Etiologic Heterogeneity

Aya Kuchiba

Department of Biostatistics, National Cancer Center

Cancer epidemiologic research typically investigates the associations between exposures and risk of a disease, in which the disease of interest is treated as a single outcome. However, most cancers, including colon cancer, breast cancer and lung cancer, are comprised of a range of heterogeneous molecular and pathologic processes, likely reflecting the influences of diverse exposures. The approach, which incorporates data on the molecular and pathologic features of a disease directly into epidemiologic studies, has been increasingly recognized to better identify causal factors and better understand how potential etiologic factors influence disease development. This paper introduces the conceptual framework and methodological development for investigating etiologic heterogeneity.

Key words: Etiologic heterogeneity, subtypes, competing risks, clustering, molecular epidemiology, cancer genomics.